Linear Regression

Dr. Jarad Niemi

STAT 4610X - Iowa State University

February 4, 2025

Outline

- Simple Linear Regression (SLR)
    - Model
    - Interpretation
    - Assumptions
    - Diagnostics
    - Example

Simple Linear Regression

For observation i, let

- Y_i be the response variable and
- X_i be the explanatory variable.

The simple linear regression model (SLR) assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

or, equivalently,

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2).$$

Interpretation

Recall

$$E[Y_i] = \beta_0 + \beta_1 X_i$$

Thus,

- β_0 is the expected response when $X_i = 0$
- β_1 is the expected increase in the response when X_i is increased by 1.

Assumptions

Recall

$$E[Y_i] = \beta_0 + \beta_1 X_i, \quad \epsilon_i \stackrel{ind}{\sim} N(0, \sigma^2)$$

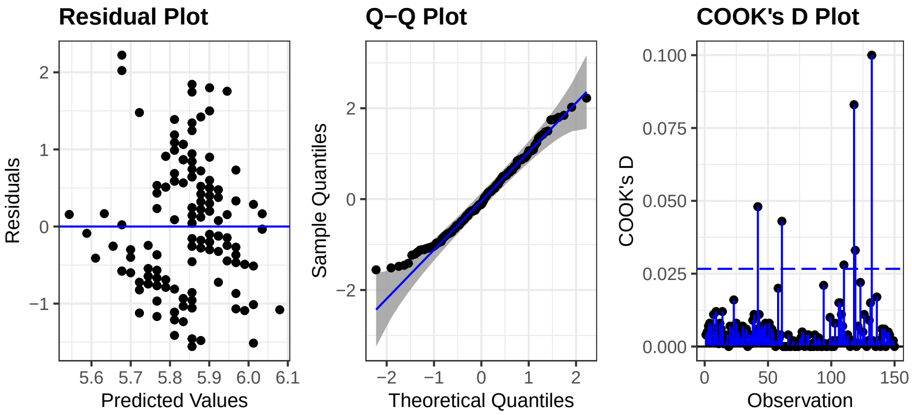
Thus, the model assumptions are

- The errors are normally distributed.
- The errors have constant variance.
- The errors are independent.
- The relationship between the expected response and the explanatory variable is a straight line.

Diagnostics

To evaluate these model assumptions we utilize diagnostic plots:

```r
m <- lm(Sepal.Length ~ Sepal.Width, data = iris)
ggResidpanel::resid_panel(m, plots = c("resid", "qq", "cookd"), qqbands = TRUE, nrow = 1)
```



Triathlon Data

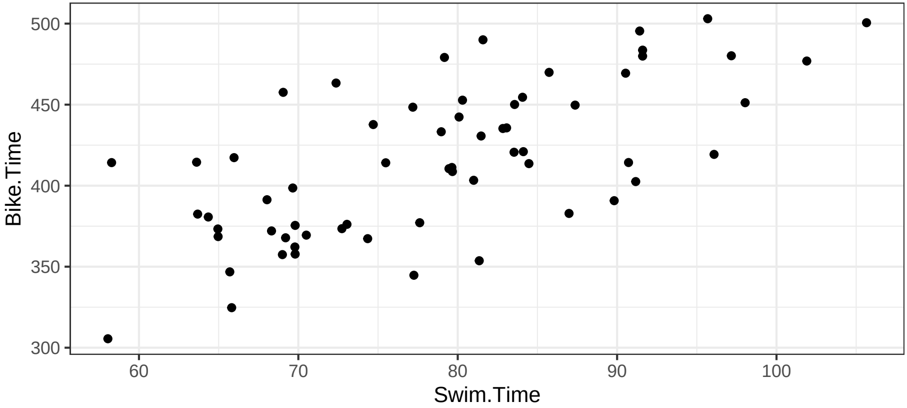
from `https://modules.scorenetwork.org/triathlons/ironman-lakeplacid-mlr/`

```
d <- read_csv("ironman_lake_placid_female_2022_canadian.csv")
head(d)

# A tibble: 6 x 17
    Bib Name     Country Gender Division Division.Rank Overall.Time Overall.Rank Swim.Time Swim.Rank Bike
  <dbl> <chr>    <chr>   <chr>  <chr>            <dbl>        <dbl>        <dbl>     <dbl>     <dbl>
1     2 Melanie~ Canada  Female FPRO                 5         575.           21      58.0        57
2     9 Pamela-~ Canada  Female FPRO                10         610.           51      65.8       253
3  1000 Carley ~ Canada  Female F35-39               4         660.          126      65.7       249
4  1935 Seanna ~ Canada  Female F45-49               3         665.          131      74.4       727
5   511 Marie-C~ Canada  Female F45-49               4         679.          161      77.2       899
6  1240 Julie H~ Canada  Female F40-44               6         693.          202      77.6       921
# i 5 more variables: Run.Time <dbl>, Run.Rank <dbl>, Finish.Status <chr>, Location <chr>, Year <dbl>
```

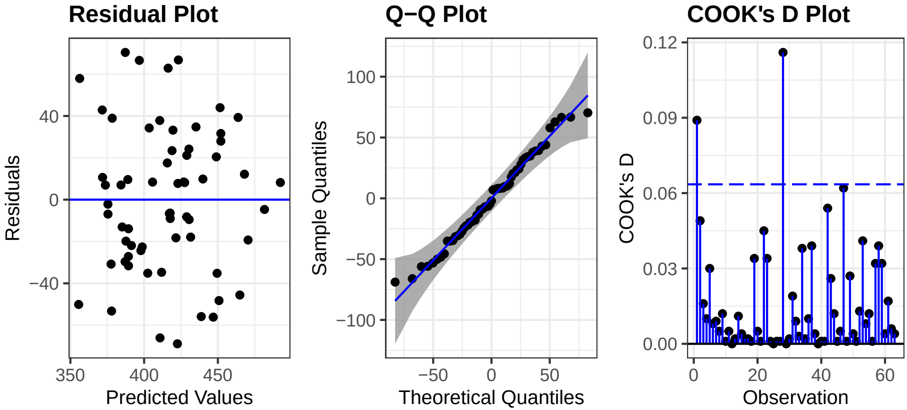
Bike Time v Swim Time

```
ggplot(d |> filter(Swim.Time < 500), aes(x = Swim.Time, y = Bike.Time)) + geom_point()
```



Bike Time v Swim Time - Model Diagnostics

```
m <- lm(Bike.Time ~ Swim.Time, data = d |> filter(Swim.Time < 500))
ggResidpanel::resid_panel(m, plots = c("resid", "qq", "cookd"), qqbands = TRUE, nrow = 1)
```



Bike Time v Swim Time - Model Results

```
summary(m)


Call:
lm(formula = Bike.Time ~ Swim.Time, data = filter(d, Swim.Time <
    500))

Residuals:
    Min      1Q  Median      3Q     Max 
-68.901 -23.468  -2.169  23.808  70.369 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 188.8604    32.0893   5.885 1.82e-07 ***
Swim.Time     2.8729     0.4035   7.120 1.44e-09 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.83 on 61 degrees of freedom
Multiple R-squared:  0.4538,Adjusted R-squared:  0.4449 
F-statistic: 50.69 on 1 and 61 DF,  p-value: 1.443e-09
```

Bike Time v Swim Time - Written Results

```
cbind(coef(m), confint(m))

                               2.5 %      97.5 %
(Intercept) 188.860386 124.693942 253.026829
Swim.Time     2.872855   2.065987   3.679724

summary(m)$r.squared

[1] 0.4538433
```

When swim time is 0, the expected Bike Time is 189 mins with a 95% interval of (125, 253).
For additional minute of swim time, the bike time is expected to increase 2.9 mins (2.1, 3.7).
The model explains 45% of the variability in bike time.

Bike Time v Swim Time - Plot

```
ggplot(d |> filter(Swim.Time < 500), aes(x = Swim.Time, y = Bike.Time)) +
  geom_point() + geom_smooth(method = "lm")
```